Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 3179 – 3183

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Mining Multi-Patterns in Pattern-Based Clustering

Qian Ma[a], Jingfeng Guo[a]*

[a]*Dept. of Information Science and Engineering, Yanshan University, QinHuangdao, Hebei ,China*

Abstract

Unlike traditional clustering methods that focus on grouping objects with similar values on a set of dimensions, pattern-based clustering finds objects that exhibit coherent patterns in subspaces. Pattern-based clustering extends the concept of traditional clustering and benefits a wide range of applications. However, most of previous approaches based on single pattern model can only explore one of specific patterns, not both of them. This paper analyses different kinds of patterns between items, presents the conception of multi-pattern model. Based on this model, we can capture patterns of shifting, scaling, and other patterns with the same feathers simultaneously. From the study of multi-pattern model's characters and operating principles, an effective algorithm is introduced to cluster objects which are coherent with multi-patterns.

Keywords: Pattern-based Clustering; Bicluster; Relation; Multi-pattern

1. Introduction

Pattern-based clustering which can be seen as an extension work of subspace clustering and biclustering[1,2,3], identifies groups of objects that show similar activity patterns under a specific subset of the columns. Since similarity between objects and attributes are based on algebraical scores, we can obtain certain clustering results.

In previous studies, the descriptions of *pattern* in different models are various[4]. Determining similarities of patterns in different models is of different ways. In most models, pattern is defined as a certain consistency of shifting or scaling which objects express in a set of attributes. To summarize the

* Corresponding author. Tel.: +86-13833569013; fax: +0-000-000-0000 .
E-mail address: jm8281@163.com.

doi:10.1016/j.proeng.2012.01.462

characteristics of pattern, it can be seen as a mode of consistency in the conduct of a trend which is a set of relation mapping, contains a wide range of content.

Consistency can be binary relation or multivariate relation. Current researches mainly focus on binary relation models, such as shifting model, scaling model, shifting-and-scaling model, and ordered model. In shifting model, consistency can be seen as $\{y=x+k\}$, (k is a const value); in scaling model as $\{y=x\times k\}$; and in ordered model as $\{y\leq x$ or $y\geq x\}$, etc. Those clustering models which can only capture single type of patterns, can not mining those pattern simultaneously.

There are three objects in 5 attributes as Table 1 shows. Based on shifting pattern model, we can get a pcluster$\{(o_1,o_2,o_3)(c_1,c_2,c_3)\}$. And based on scaling pattern model, pcluster$\{(o_1,o_2,o_3)(c_3,c_4,c_5)\}$ can be found.

Table 1. An example of Expression data with three objects and five columns

	c_1	c_2	c_3	c_4	c_5
o_1	1	11	2	8	4
o_2	3	13	4	16	6
o_3	4	14	5	20	8

In fact, those two patterns could be combined into one cluster based on multi-pattern which is more meaningful, as Figure 1 shows, but in current single-pattern models this is without considering.

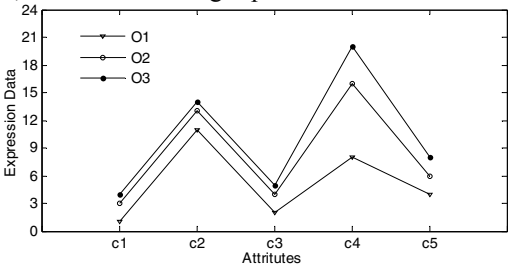

Figure 1. Cluster based on multi-pattern of data in Table 1

In real database, various relevant patterns will exist simultaneously in the objects and attributes, for performance-related trends or relevant functions. How to correctly assess the consistency of information is an important issue.

2. Related work and our contribution

There are numerous pattern-based clustering models due to various definitions of patterns and coherence measures. Shifting and scaling patterns are most widely studied in nowadays，and there are a lot of mature algorithms. Cheng et al. put forward of Bicluster which is the predecessor of pCluster[5]. Bicluster mainly aims on finding out the genes with consistent in a DNA microarray. They measured the coherence of clusters by the mean squared residue score. Wang et al. proposed the concept of pattern-based Clustering which introduced the notion of pScore to measure the similarity between the objects in clusters[6,7]. Effective data exploration can focus on the shifting/scaling patterns[8]. Xin et al. introduce reg-cluster algorithm to mining shifting-and-scaling patterns[9]. And the ordered pattern model has been proposed by Wang et al based on the sequence of the clustering method seqClu[10].

Based on the previous research, this paper focuses on the study of clustering methods based on multi-pattern model. Innovative contributions have been made in following aspects.

- This paper analyses different kinds of patterns between objects and attributes, make general theoretical analysis of pattern diversity.
- Based on theoretical study, we analyze the characteristics and computing principles of different patterns.
- A new model of multi-pattern based clustering is proposed, as well as the algorithm which can capture multi-patterns simultaneously. Experimental result shows that our algorithm is effective.

3. Model of Multi-pattern

3.1. Relation and pattern

Definition 1: Relation. Let a and b be two objects, $C=\{c_1,c_2,...,c_i,\ldots,c_n\}$ be a subset of attributes. Expression data vectors of the two objects respectively be

$$D_a = \{d_{a1}, d_{a2}, ..., d_{ai}, ..., d_{an}\} \quad (1)$$

and

$$D_b = \{d_{b1}, d_{b2}, ..., d_{bi}, ..., d_{bn}\} \quad (2)$$

Let R be a set of ordered pair. R is called a Relation form a to b if for each $i\in[1,n]$:

$$< d_{ai}, d_{bi} > \in R \quad (3)$$

a is said to be related to b by R. Denoted as $b=R(a)$.

Relation can also been defined between attributes as well as objects, which represents relationships between elements of two or more attributes. Relation can be seen as a set of ordered pairs denoted binary relation, and which current pattern-based clustering studies focus on. There are several properties that are used to classify relations on a set; the most important are reflexivity, symmetry and transitivity.

Let $O=\{o_1,o_2,\ldots,o_m\}$ be a subset of objects and $C=\{c_1,c_2,\ldots,c_n\}$ be a subset of attributes, d_{ij} be expression data object o_i express in attribute c_j, R be the relation on objects set O.

- Reflexivity: Relation R on objects set O called reflexive if for each $d_{ij}=d_{kj}$, $<d_{ij},d_{kj}>\in R$, for any $c_j\in C$.
- Symmetry: Relation R on objects set O is called symmetric if $r_1<d_{ij},d_{kj}>\in R$, whenever $r_2<d_{kj},d_{ij}>\in R$, for any $c_j\in C$, r_1 is inverse function of r_2. Relation R such that $<d_{ij},d_{kj}>\in R$ and $<d_{kj},d_{ij}>\in R$ only if $d_{ij}=d_{kj}$, for any $c_j\in C$, is called antisymmetric.
- Transitivity: A relation R on objects set O is called transitive if whenever $r_1<d_{ij},d_{kj}>\in R$ and $r_2<d_{kj},d_{hj}>\in R$, then $r_3<d_{ij},d_{hj}>\in R$, for any $c_j\in C$. r_3 is the composite function of r_1 and r_2.

Lemma 1: Let $r_1=<s_1,e_1>$ and $r_2=<s_2,e_2>$ be two ordered pair in relation set. Use the notation $r_3=r_1\circ r_2$ to indicate that r_3 is the composite relation of r_1 and r_2, then:

$$r_3 = r_1 \circ r_2 = \begin{cases} \phi & e_1 \neq s_2 \\ < s_1, e_2 > & e_1 = s_2 \end{cases} \quad (4)$$

Definition 2: Pattern. Set D as dataset for a collection of objects, each object has been described by a set of attributes A. $O\in D$ for a subset of objects, $T\in A$ for a subset of attributes. Tuple (O, T) is a matrix. If each pair of object of O in T has relation $r\in R$, then this consistency of the performance is called a *pattern* and it is recorded as P.

To sum up, a complete pattern should include three important elements: the object set O, attribute set T, and coherence relation set R. Therefore, the formal pattern can be described as: P (O, T, R).

3.2. Order of pattern

Let R_1 and R_2 be two relations, R_1 is said to be included in R_2 if and only if every ordered pair<a, b> of R_1 is also an element of R_2. We use the notation $R_1 \subseteq R_2$ to indicate that R_1 is a sub-relation of R_2.

This relation of inclusion widely exists. Relation of identity pattern is included in shifting or scaling pattern, relation of shifting/scaling pattern is included in shifting-and-scaling pattern, and both of them are sub-relation of relation in linear pattern. Different types of relations have different influences on quality of clusters, and there will be much overlapping between patterns with inclusive relations. Based on inclusion of relations, we give different weight value to patterns; the weight value is called Order. Basic Order is defined by users; the simpler pattern model would be given smaller Basic Order value.

Definition 3: Order of pattern. Set D as dataset for a collection of objects, each object has been described by a set of attributes A. $O \in D$ for a subset of objects, $T \in A$ for a subset of attributes. Tuple (O, T) is a matrix. $P(O,T,R)$ describe a pattern with relation of Basic Order k. Order of P is:

$$Order(P) = \frac{k \times |O| |T|}{|D| |A|} \tag{7}$$

Let $\{R_1,R_2,\ldots,R_n\}$ be relation sets of patterns which considered in multi-pattern model. We have $Order(R_i)=Order(R_j)$ when $R_i=R_j$; $Order(R_i)<Order(R_j)$ when $R_i \subset R_j$.

Let $P_1(O_1,T_1,R_1)$ and $P_2(O_2,T_2,R_2)$ be two patterns, (O_1,T_1) is said to be more coherent than (O_2,T_2) if $Order(P_1)<Order(P_2)$.

Lemma 2: Let R_1 and R_2 be two relations, use $R_3=R_1 \infty R_2$ to indicate that R_3 is the combination relation of R_1 and R_2, then:

$$R_3 = R_1 \infty R_2 = R_1 \cup R_2 \cup \Big(\bigcup_{r_1 \in R_1, r_2 \in R_2} r_1 \circ r_2 \Big) \tag{8}$$

3.3. Multi-pattern-based model

In multi-pattern models, not only a single pattern, but various types of patterns are considered synthetically. How to solve different types of patterns and combine clusters with different relation has been a key problem.

Because of the properties of transitivity, when combining two patterns, all the elements in combination pattern are related to each other and the new relation set can be obtained through composite functions.

Lemma 3: Let $P_1(O_1,T_1,R_1)$ and $P_2(O_2,T_2,R_2)$ be two patterns, $O_1 \cap O_2 \neq \emptyset$, $T_1 \cap T_2 \neq \emptyset$, $P_o(O_o,T_o,R_o)$ be a pattern. Use notation $P_o=P_1 \otimes P_2$ to indicate that P_o is the combination with objects of P_1 and P_2, have:

$$\begin{aligned} &P_o = P_1 \otimes P_2 \\ &O_o = O_1 \cup O_2; \quad T_o = T_1 \cap T_2; \quad R_o = (R_1 \infty R_2) \cap (O_o^2 \cup T_o^2) \end{aligned} \tag{9}$$

and

$$Order(R_o) = \frac{|O_1| |T_1| \times Order(R_1) + |O_2| |T_2| \times Order(R_2)}{|O_o| |T_o|} \tag{10}$$

The combination within attributes of two patterns P_c can be calculated as well.

$$\begin{aligned} &P_c = P_1 \oplus P_2 \\ &O_o = O_1 \cap O_2; \quad T_o = T_1 \cup T_2; \quad R_o = (R_1 \infty R_2) \cap (O_o^2 \cup T_o^2) \end{aligned} \tag{9}$$

4. Algorithm of Multi-pattern-based clustering

Relations of cluster with lower Order will have better quality. And mining relations of lower Order should be easier and faster than those complex relations with high Order. Preference for the mining or combining of clusters with lower order relation will optimize the clustering process.

Algorithm of Multi-pattern based Clustering as follows:

- Step 1: Define proper basic order to each type of pattern which considered in our algorithm;
- Step 2: Mining candidate clusters with each type of pattern model form lower order to higher order;
- Step 3: Do combination of candidate patterns form lower order to higher, continue this process until there aren't any new patterns been found;
- Step 4: Output patterns which satisfy user defined scales.

5. Conclusion

In this paper, we make detail analysis about diversity of pattern. Patterns with different types of relation can be combined together based on definition of order in different pattern model. A multi-pattern-based clustering model has been proposed in this paper, as well as the algorithms to mining multi-pattern clusters. Several interesting and important problems still remain open, such as how to identify more meaningful patterns and relations, how to mining nonlinear patterns and covering them to multi-pattern models.

Acknowledgements

This paper is supported by a grant from the Hebei provincial Applied Basic Research Program, China, (Grant No.10963527D).

References

[1]C. C. Aggarwal and P. S. Yu. Finding generalized projected clusters in high dimensional spaces. In: SIGMOD, p. 70–81.

[2]Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In: Proceedings of the ACM SIGMOD Conference,1998, p. 94–105.

[3]C. H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In Proceedings of SIGKDD Conference, 1999, p. 84–93.

[4]Daxin Jiang, Jian Pei, Aidong Zhang.: A General Approach to Mining Quality Pattern-Based Clusters from Microarray Data. In: proc. of the 2005 DASFAA, p. 188–200.

[5]Y. Cheng and G.M. Church.: Biclustering of expression data. In: Proc. of ISMB'00, p. 93–103. AAAI Press.

[6]Wang H,Wang W,Yang J,Yu PS.Clustering by pattern similarity in large data sets. Proc. of the 2002 ACM SIGMOD, p. 394–405.

[7]Jiong Yang, Wei Wang, Haixun Wang, and Philip S. Yu.: δ-clusters: Capturing subspace correlation in a large data set. In ICDE, p. 517–528.

[8] Pei, J., Zhang, X., Cho, M., et al. MaPle: A Fast Algorithm for Maximal Pattern-based Clustering. ICDM'03.

[9] X. Xu, Y.Lu, et al. Mining Shifting-and-Scaling Co-Regulation Patterns on Gene Expression Profiles, Proc. of the 22nd International Conference on Data Engineering, 2006, p. 89 - 99.

[10] Liu, J.,Wang W, OP-Cluster: Clustering by Tendency in High Dimensional Space.In ICDM'03.